Glass Identification

- ## Introduction

The dataset chosen for the analysis was the **Glass Identification**. It was chosen since there are over 100 instances and more than ten attributes in the dataset. Aside from that, I picked this dataset because of The study of classification of types of glass was motivated by criminological investigation. At the scene of the crime, the glass left can be used as evidence...if it is correctly identified! All the observations observed during the analysis are included in the report. Specifically, we aimed to determine which factors contribute to glass type by analysing the dataset. In order for the reader to get the most out of the dataset, we have included all necessary plots and graphs. A complete analysis was conducted using R studio using the R programming language.

- ## About Dataset

Glass fragments are one of the most frequently used items in forensic science. In most of the crime scenes such as house-breaking, even small fragments of the glass attached to the clothes of the person who is suspected would solve the problem. However, we are not certain that all the input variables are relevant. Thus, it may be worthwhile to test various selection methods. The most useful characteristics of glass for forensic purposes represents the refractive index (RI) which has a high precision even for small pieces. For even larger fragments, the elemental components can be obtained using a Scanning Electron Microscope. The data collected for 214 glass samples with 10 attributes and were analysed at the Home Office Forensic Science Laboratory, Birmingham.

Attribute Information:

1. Id number: 1 to 214
2. RI: refractive index
3. Na: Sodium (unit measurement: weight percent in corresponding oxide, as are attributes 4-10)
4. Mg: Magnesium
5. Al: Aluminum
6. Si: Silicon
7. K: Potassium
8. Ca: Calcium
9. Ba: Barium
10. Fe: Iron
11. Type of glass: (class attribute)
-- 1 building_windows_float_processed
-- 2 building_windows_non_float_processed
-- 3 vehicle_windows_float_processed
-- 4 vehicle_windows_non_float_processed (none in this database)
-- 5 containers
-- 6 tableware
-- 7 headlamp

- **Data Analysis**

Using R, the datasets were imported into a data frame in csv format. When the dataset was first examined, it appeared to be loaded correctly. There are 214 observations and 10 observations in data frame. We are removing the ID column, because the columns in our dataset are named from 0 to 10 which is ambiguous and difficult to read and interpret which are not required and are not important that is "**ID"** column**.**

Here is the summarized dataset after validation:

```
       RI              Na              Mg              Al              Si               K                 Ca               Ba              Fe
 Min.   :1.511   Min.   :10.73   Min.   :0.000   Min.   :0.290   Min.   :69.81   Min.   :0.0000   Min.   : 5.430   Min.   :0.000   Min.   :0.00000
 1st Qu.:1.517   1st Qu.:12.91   1st Qu.:2.115   1st Qu.:1.190   1st Qu.:72.28   1st Qu.:0.1225   1st Qu.: 8.240   1st Qu.:0.000   1st Qu.:0.00000
 Median :1.518   Median :13.30   Median :3.480   Median :1.360   Median :72.79   Median :0.5550   Median : 8.600   Median :0.000   Median :0.00000
 Mean   :1.518   Mean   :13.41   Mean   :2.685   Mean   :1.445   Mean   :72.65   Mean   :0.4971   Mean   : 8.957   Mean   :0.175   Mean   :0.05701
 3rd Qu.:1.519   3rd Qu.:13.82   3rd Qu.:3.600   3rd Qu.:1.630   3rd Qu.:73.09   3rd Qu.:0.6100   3rd Qu.: 9.172   3rd Qu.:0.000   3rd Qu.:0.10000
 Max.   :1.534   Max.   :17.38   Max.   :4.490   Max.   :3.500   Max.   :75.41   Max.   :6.2100   Max.   :16.190   Max.   :3.150   Max.   :0.51000
 glass_type
 1:70
 2:76
 3:17
 5:13
 6: 9
 7:29
```

From the summary, it shows that the dataset is not normally distributed and also, we can see that mean of SI is larger which is 72.65 which is much bigger than other. We have to perform some log transform and also have to check some outliers to clean the data. To make the more clear and correct result we have to perform some data cleaning process on the data frame to make the data more normalize and need to remove the outliers. First, we will perform the log transform to the SI to make it more normalise for the dataset. After clearing the outliers and performing the log transform on SI and also normalising the dataset we will get a new dataset which we will use. However, to make it more clear rather than using log transform we will use min max normalization to make our date more normalise for the further classification and will just remove the outliers. Below is the summary of normalise data set after removing the outliers.

```
       RI               Na               Mg               Al               Si               K                  Ca               Ba
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.4056   1st Qu.:0.4007   1st Qu.:0.5356   1st Qu.:0.2827   1st Qu.:0.4627   1st Qu.:0.02375   1st Qu.:0.3957   1st Qu.:0.0000
 Median :0.4941   Median :0.4779   Median :0.7762   Median :0.3396   Median :0.5520   Median :0.09018   Median :0.4463   Median :0.0000
 Mean   :0.5212   Mean   :0.5056   Mean   :0.6197   Mean   :0.3662   Mean   :0.5286   Mean   :0.08250   Mean   :0.4720   Mean   :0.0584
 3rd Qu.:0.6001   3rd Qu.:0.5903   3rd Qu.:0.8040   3rd Qu.:0.4174   3rd Qu.:0.6054   3rd Qu.:0.09823   3rd Qu.:0.5149   3rd Qu.:0.0000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
       Fe            glass_type
 Min.   :0.0000   Min.   :1.000
 1st Qu.:0.0000   1st Qu.:1.000
 Median :0.0000   Median :2.000
 Mean   :0.1071   Mean   :2.804
 3rd Qu.:0.1765   3rd Qu.:3.000
 Max.   :1.0000   Max.   :7.000
```

Here is the first 10 observations of the new dataset:

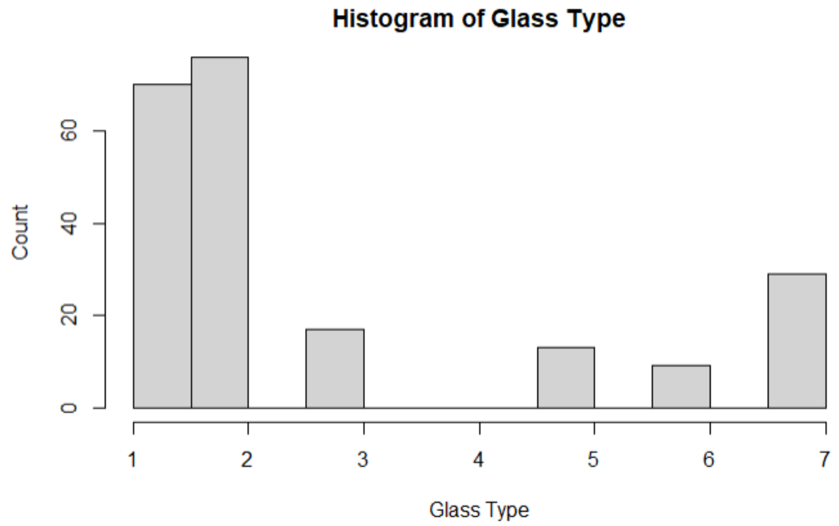
Description: df [6 x 10]

	RI <dbl>	Na <dbl>	Mg <dbl>	Al <dbl>	Si <dbl>	K <dbl>	Ca <dbl>	Ba <dbl>	Fe <dbl>
1	0.7584050	0.5483193	1.0000000	0.2523364	0.3572779	0.009661836	0.4695898	0	0.0000000
2	0.4925723	0.6008403	0.8017817	0.3333333	0.5368620	0.077294686	0.3394625	0	0.0000000
3	0.3807662	0.5252101	0.7906459	0.3894081	0.5860113	0.062801932	0.3323904	0	0.0000000
4	0.4964816	0.4579832	0.8218263	0.3115265	0.5141777	0.091787440	0.3946252	0	0.0000000
5	0.4777170	0.4705882	0.8062361	0.2959502	0.6030246	0.088566828	0.3734088	0	0.0000000
6	0.3635653	0.3697479	0.8040089	0.4143302	0.5822306	0.103059581	0.3734088	0	0.5098039

6 rows | 1-10 of 10 columns

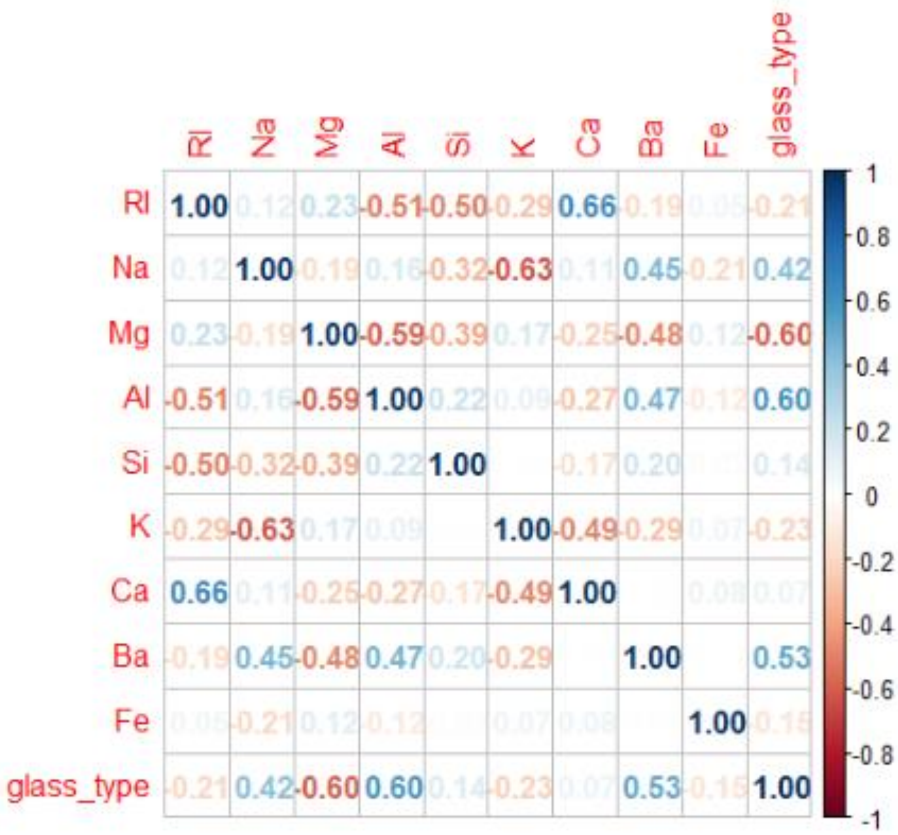
- **Data Visualisation**

1. Histogram



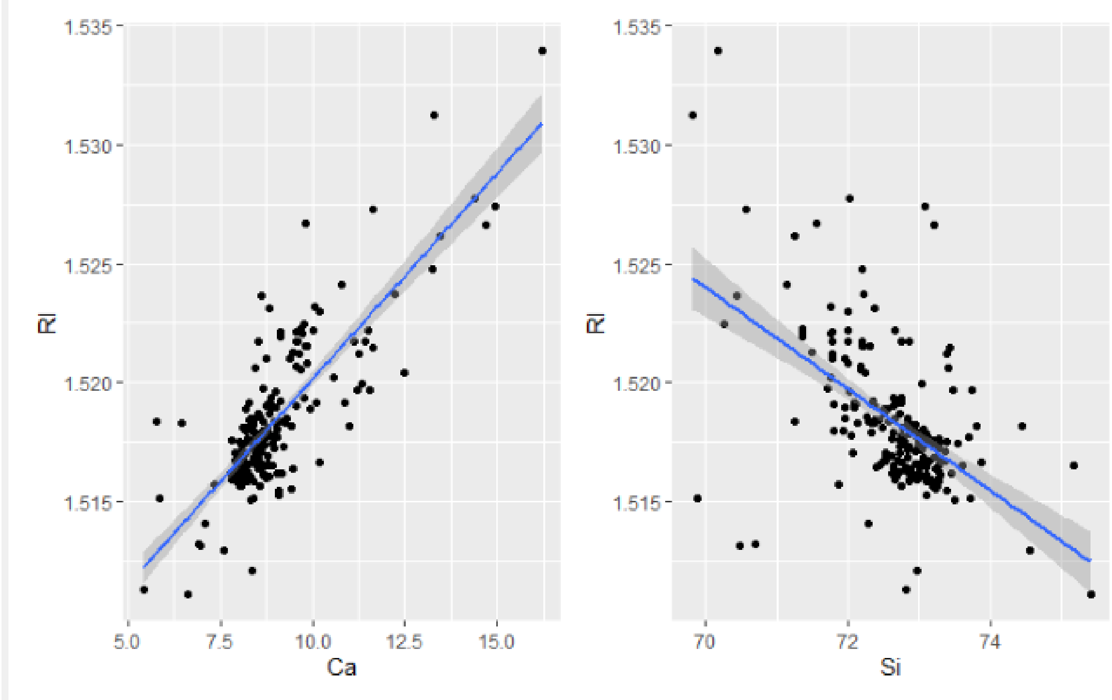
From the above histogram, we can see that the data is not normally distributed as the glass type of 1 and 2 means building_windows_float_processed and building_windows_non_float_processed respectively higher than other. As the glass type 1 and 2 consist of more than 67% of the glass types.

2. Correlation Plot



From the above correlation plot, we can see that the RI is highly correlated with Ca compare to other as it is 0.66. Similarly, Al and Ba is also correlated to each with 0.47 and also Ba and Na is correlated to each other by 0.41. Also, with this we can see that, Na, Al, Ba are some of the variables which is correlated with the glass type. To come with the better conclusion, we will plot correlation coefficient plot of the dataset.

3. Correlation Coefficient:



It is evident from the figure that the Refractive Index and Calcium have a strong linear relationship as the correlation coefficient is 0.66 which suggests these two variables are highly correlated.

Furthermore, it appears that the refractive index of the glass decreases with increasing silicon (i.e. - 0.50), indicating that an increase in Silicon decreases its refractive index.

In the conclusion, Oxides of calcium and silicon are the best predictors of refractive index and Mg and Al are the best predictors for glass type.

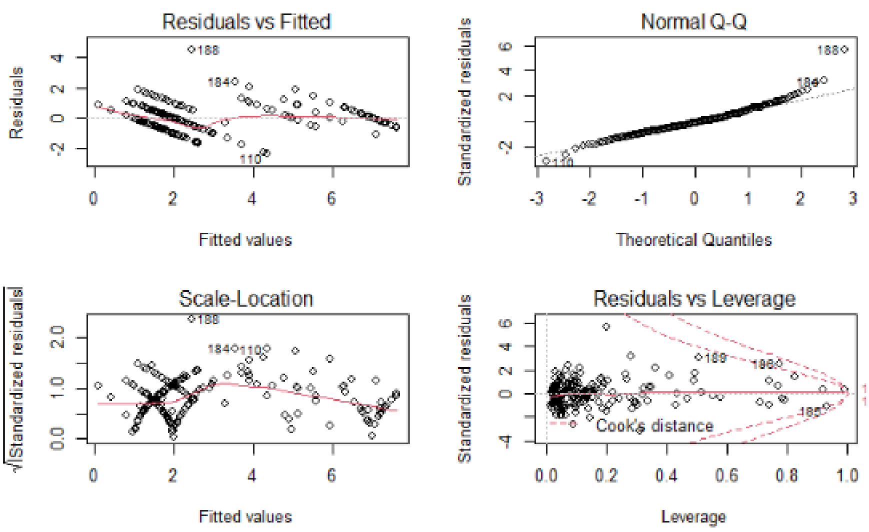
4. Linear Regression and Four-panel plot :

```
Call:
lm(formula = glass_type ~ RI + Na + Mg + Al + Si + K + Ca + Ba +
    Fe + RI:Na + RI:Mg + RI:Al + RI:Si + RI:K + RI:Ca + RI:Fe +
    Na:Mg + Na:Al + Na:K + Na:Ba + Na:Fe + Mg:Ca + Mg:Ba + Mg:Fe +
    Al:K + Al:Fe + Si:K + Si:Fe + K:Ca + K:Fe + Ca:Fe, data = glass)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3518 -0.5130 -0.0614  0.4342  4.5330

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.256e+04  1.068e+04   3.049 0.002641 **
RI          -2.147e+04  7.045e+03  -3.047 0.002652 **
Na          -4.290e+02  1.196e+02  -3.587 0.000429 ***
Mg          -3.089e+02  1.004e+02  -3.077 0.002414 **
Al          -7.075e+02  1.712e+02  -4.133 5.45e-05 ***
Si          -2.941e+02  1.108e+02  -2.655 0.008644 **
K           -7.020e+02  2.254e+02  -3.115 0.002140 **
Ca          -3.673e+02  1.048e+02  -3.505 0.000575 ***
Ba          -1.609e+01  3.779e+00  -4.256 3.32e-05 ***
Fe           1.056e+02  1.114e+03   0.095 0.924641
RI:Na        2.827e+02  7.885e+01   3.585 0.000433 ***
RI:Mg        1.954e+02  6.621e+01   2.951 0.003584 **
RI:Al        4.718e+02  1.122e+02   4.206 4.07e-05 ***
RI:Si        1.940e+02  7.307e+01   2.655 0.008631 **
RI:K         5.259e+02  1.527e+02   3.445 0.000709 ***
RI:Ca        2.418e+02  6.902e+01   3.503 0.000579 ***
RI:Fe        1.018e+03  6.041e+02   1.686 0.093533 .
Na:Mg        5.995e-01  1.008e-01   5.945 1.38e-08 ***
Na:Al       -4.685e-01  2.504e-01  -1.871 0.062928 .
Na:K        -1.469e+00  4.025e-01  -3.650 0.000343 ***
Na:Ba        1.090e+00  2.628e-01   4.148 5.14e-05 ***
Na:Fe       -2.023e+01  5.268e+00  -3.840 0.000169 ***
Mg:Ca        3.564e-01  8.923e-02   3.995 9.40e-05 ***
Mg:Ba        4.517e-01  1.662e-01   2.718 0.007206 **
Mg:Fe       -1.239e+01  4.166e+00  -2.974 0.003339 **
Al:K        -2.474e+00  6.981e-01  -3.544 0.000500 ***
Al:Fe       -1.617e+01  5.907e+00  -2.737 0.006814 **
Si:K        -8.205e-01  3.133e-01  -2.619 0.009573 **
Si:Fe       -1.586e+01  4.721e+00  -3.360 0.000951 ***
K:Ca        -1.577e+00  3.381e-01  -4.666 5.95e-06 ***
K:Fe        -2.671e+01  7.990e+00  -3.343 0.001007 **
Ca:Fe       -1.803e+01  4.424e+00  -4.075 6.87e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the linear regression, we can see that for detecting the glass type all factors are important however, FE is not that much important it is not significant to the glass type. Adding to the point, we can see that interaction term of RI and Al is highly significant and states that this interacting term might be more important to detect the glass type also interaction term with Na and Mg shows the same relation and with the highly significant which has smaller p-value than other and help for indicating the glass type. However, the interacting term like Na:Al, RI:Fe, are not that much important term for detecting the glass type. Here is the four-panel plot for the model.



Code :

```
```{r}
library(ggplot2)
require(GGally)
library(gridExtra)
library(reshape)
# Glass Identifiction

glass = read.csv('glass.data', header = FALSE, as.is = FALSE)
names = c('Id','RI','Na','Mg','Al','Si','K','Ca','Ba','Fe','glass_type')
names(glass) = names
glass = subset(glass, select = c(-1) )


a = outlier(glass, plot = FALSE)
out <- boxplot(glass$RI, range = 2)$out
out_ind <- which(glass$RI %in% c(out))
out_ind

glass[out_ind, ]
glass1<- glass[-which(glass$RI %in% c(out)),]
boxplot(glass1)

min_max_norm <- function(x) {
    (x - min(x)) / (max(x) - min(x))
}
glass_norm <- as.data.frame(lapply(glass1[1:9], min_max_norm))
glass_norm$glass_type = glass1$glass_type
summary(glass_norm)
summary(glass1)
head(glass_norm)
```
```

```
```{r}
hist(glass$glass_type, main = "Histogram of Glass Type", xlab = "Glass Type", ylab= "Count")
```
```{r}
library(corrplot)
str(glass)

E2 = cor(glass_norm, method = 'spearman')
corrplot(E2, method="number")
```
```{r}
ggpairs(glass_norm)
```
```
```

```
```{r}
plot1<-ggplot(glass,aes(x = Ca, y = RI)) + geom_point() +geom_smooth(method = "lm")
plot2<-ggplot(glass,aes(x = Si, y = RI)) + geom_point() +geom_smooth(method = "lm")
grid.arrange(plot1, plot2, ncol=2)
```
```{r}
model = lm(glass_type~.*., data = glass)
smodel = step(model, trace = FALSE)
summary(smodel)
par(mfrow =c(2,2))
plot(smodel)
```
```
```

El Nino

- ## Introduction

El Nino was the dataset chosen for the analysis. This cycle of El Nino/Southern Oscillation (ENSO) of 1982-1983, the strongest of the century, brought many problems to various parts of the world. The western Pacific regions were hit by drought and devastating brush fires, while Peru and the United States had destructive floods from increased rainfalls. No one could have predicted or detected the ENSO cycle until it was near its peak. In order to study ocean-atmosphere interactions in large scale on seasonal to interannual timescales, an ocean observing system was needed (such as the TAO array).

Globally, the TAO array gives scientists, weather prediction centres, and climate researchers access to real-time data. Using ENSO cycle data, tropical Pacific Ocean temperatures can be forecast one or two years in advance. We can forecast the weather based on moored buoys, drifting buoys, volunteer ship temperature probes, and sea level measurements.

According to Wikipedia, “El Nio–Southern Oscillation (ENSO) is an irregularly periodic variation of winds and sea surface temperatures over the tropical eastern Pacific Ocean that affects a large area of tropical and subtropical climates.

- ## About Dataset

The variables included in the data are: date, latitude, longitude, zonal winds (west *0, east>0), meridian winds (south *0, north>0), humidity, air temperature, sea surface temperature, and subsurface temperatures down to a depth of 500 meters. Some buoys have data from as far back as 1980. Rainfall, solar radiation, current levels, and subsurface temperatures were also recorded at various locations.

The dataset contained a lot of missing data, mostly in a Missing Not At Random pattern. Due to the fact that the buoys were commissioned at different times of the year, the amount of data collected varies. A range of 18 years separates the year of launching of the buoys, from 1980 to 1998. Additionally, the amount of data available is also dependent on the buoys' reliability.

A total of 14 percent of observations were not included in "Zonal Winds" and "Meridional Winds," 37 percent of observations were not included in "Humidity," and 10 percent of observations were not included in "Air Temp" and "Sea Surface Temp.". These missing values are represented by ".". After using the "Replace" function, we replaced these missing values with "null," removing most of the rows with null values. In the end, the size of the resulting dataset was 60% of the original.

- ## Data Analysis

Initially, the dataset was supposed to provide forecasts of tropical Pacific Ocean temperatures and El Nino events over the next 1 to 2 years. However, this process requires ad hoc knowledge in environmental science, which is outside of our areas of expertise. We are now set to investigate the relationship among variables and years rather than working on a project beyond our capabilities, which would require a lot of additional data.

Using R, the datasets were imported into a data frame in csv format. Then after transformed the missing values with null and removing the same. Also labelling the columns by checking the column names from the tao-all2.col from the UCI. We had new dataset of El Nino original dataset which is just 60% of the original.

Here is the summarized dataset after validation:

```
      obs             year           month             day              date           latitude          longtitude
 Min.   :  4060   Min.   :89.00   Min.   : 1.000   Min.   : 1.00   Min.   :891129   Min.   :-8.3300   Min.   :-180.00
 1st Qu.: 54324   1st Qu.:93.00   1st Qu.: 3.000   1st Qu.: 8.00   1st Qu.:931028   1st Qu.:-2.1600   1st Qu.:-155.00
 Median : 98083   Median :95.00   Median : 6.000   Median :16.00   Median :950430   Median : 0.0100   Median :-125.00
 Mean   : 95284   Mean   :94.83   Mean   : 6.501   Mean   :15.74   Mean   :948931   Mean   : 0.3048   Mean   : -70.84
 3rd Qu.:137050   3rd Qu.:96.00   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:961205   3rd Qu.: 4.9800   3rd Qu.: -94.96
 Max.   :178079   Max.   :98.00   Max.   :12.000   Max.   :31.00   Max.   :980623   Max.   : 9.0500   Max.   : 170.01
   zon.winds          mer.winds            humidity        air.temp        ss.temp
 Min.   :-10.700   Min.   :-10.60000   Min.   :52.10   Min.   :17.54   Min.   :18.19
 1st Qu.: -5.900   1st Qu.: -2.10000   1st Qu.:77.70   1st Qu.:26.35   1st Qu.:27.05
 Median : -4.100   Median : -0.10000   Median :81.30   Median :27.46   Median :28.37
 Mean   : -3.353   Mean   : -0.04646   Mean   :81.33   Mean   :27.06   Mean   :27.88
 3rd Qu.: -1.500   3rd Qu.:  2.00000   3rd Qu.:84.80   3rd Qu.:28.21   3rd Qu.:29.22
 Max.   : 14.300   Max.   : 13.00000   Max.   :99.90   Max.   :31.48   Max.   :31.04
```

From the summary, it shows that the dataset is correctly imported and also, we can see that year 1994 contains the most observations (15761). We will use the year 1994 to for analysis as it contains the most observation along with this year 1996 also has second most observation however as year 1994 is the most so we will use that one for our linear regression. After putting the year 1994 observation to a new dataset, here is the first 6 observations and summary of the dataset:

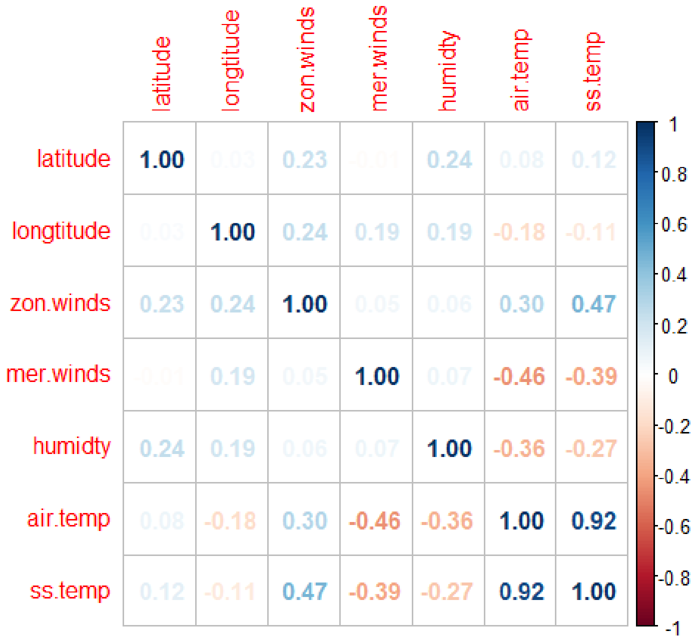
```
     year        month            day           latitude          longtitude
zon.winds        mer.winds
 Min.   :94   Min.   : 1.00   Min.   : 1.00   Min.   :-8.3100   Min.   :-179.99   Min.
:-10.40   Min.   :-10.0000
 1st Qu.:94   1st Qu.: 4.00   1st Qu.: 8.00   1st Qu.:-2.1800   1st Qu.:-155.01   1st
Qu.: -5.80   1st Qu.: -1.6000
 Median :94   Median : 7.00   Median :16.00   Median : 0.0000   Median :-124.91
Median : -4.10   Median :  0.3000
 Mean   :94   Mean   : 6.81   Mean   :15.79   Mean   : 0.1611   Mean   : -75.01   Mean
: -3.36   Mean   :  0.3618
 3rd Qu.:94   3rd Qu.:10.00   3rd Qu.:23.00   3rd Qu.: 4.9900   3rd Qu.: -94.99   3rd
Qu.: -1.50   3rd Qu.:  2.4000
 Max.   :94   Max.   :12.00   Max.   :31.00   Max.   : 9.0300   Max.   : 165.20   Max.
: 14.30   Max.   :  9.6000
    humidity        air.temp        ss.temp
 Min.   :56.00   Min.   :18.69   Min.   :18.76
 1st Qu.:77.50   1st Qu.:25.94   1st Qu.:26.67
 Median :81.30   Median :27.29   Median :28.18
 Mean   :81.41   Mean   :26.88   Mean   :27.68
 3rd Qu.:85.20   3rd Qu.:28.23   3rd Qu.:29.29
 Max.   :99.60   Max.   :30.31   Max.   :30.97
```

	year <int>	month <int>	day <int>	latitude <dbl>	longtitude <dbl>	zon.winds <dbl>	mer.winds <dbl>	humidity <dbl>	air.temp <dbl>
4297	94	1	1	-0.01	-109.99	-4.3	2.6	89.9	23.21
4298	94	1	2	-0.01	-109.99	-4.1	1.0	90.0	23.16
4299	94	1	3	-0.01	-109.99	-3.0	1.6	87.7	23.14
4300	94	1	4	0.00	-110.00	-3.0	2.9	85.8	23.39
4301	94	1	5	-0.01	-109.99	-3.4	2.0	87.8	23.53
4302	94	1	6	-0.01	-109.98	-3.2	3.1	87.2	23.71

6 rows | 1-10 of 10 columns

- **Data Visualisation**

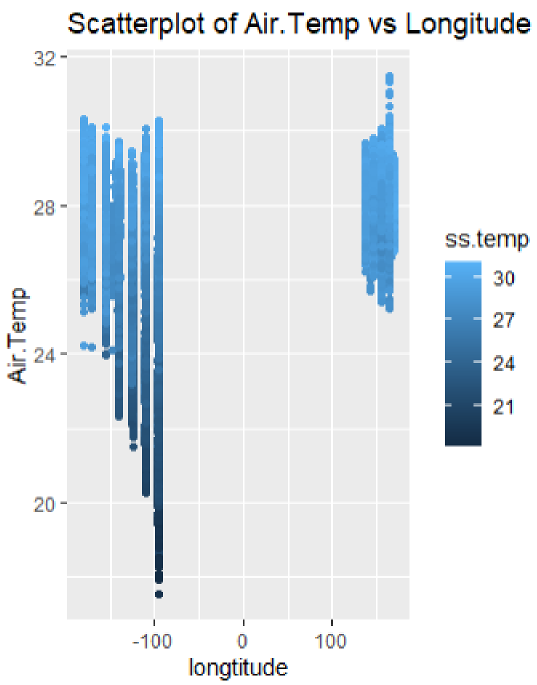
1. **Correlation Plot**

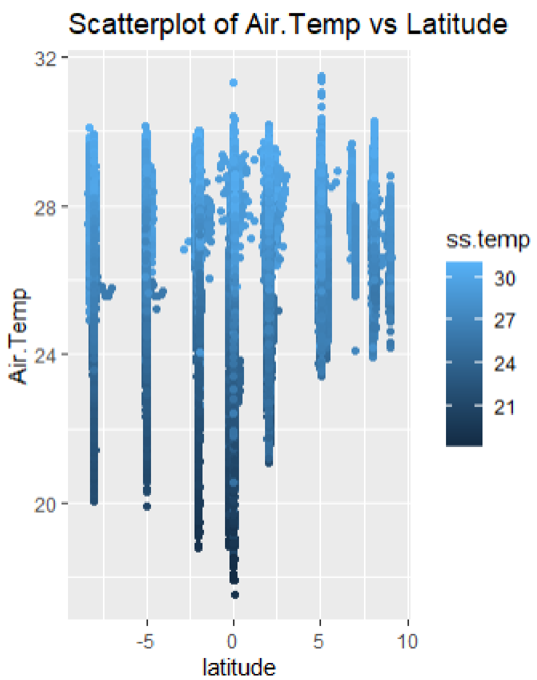


	latitude	longtitude	zon.winds	mer.winds	humidty	air.temp	ss.temp
latitude	1.00	0.03	0.23	-0.01	0.24	0.08	0.12
longtitude	0.03	1.00	0.24	0.19	0.19	-0.18	-0.11
zon.winds	0.23	0.24	1.00	0.05	0.06	0.30	0.47
mer.winds	-0.01	0.19	0.05	1.00	0.07	-0.46	-0.39
humidty	0.24	0.19	0.06	0.07	1.00	-0.36	-0.27
air.temp	0.08	-0.18	0.30	-0.46	-0.36	1.00	0.92
ss.temp	0.12	-0.11	0.47	-0.39	-0.27	0.92	1.00

This section focuses on examining the correlations among independent variables for year 1994. We would need the results to find out whether we would encounter multicollinearity during the regression analysis. As expected, sea surface temperatures are highly correlated with air temperatures. Because the former is the dependent variable, a high correlation between the two shouldn't pose a problem. Similarly, zonal wind and sea surface temperatures are correlated. The correlation is nearly 0.5.

2. **Scatter Plot**





As the Air Temperation and Sea surface are highly correlated to each other, to get clear view of both. We are plotting them as scatter plot and we can see that where the air temperature is low the surface temperature is low. However, where the Air temperature is high the surface temperature is getting high with respective to it and showing the highly correlation with each other.

3. Linear Regression

```
Call:
lm(formula = ss.temp ~ . * ., data = year941)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.9833 -0.3453 -0.0490  0.2984  3.3290 

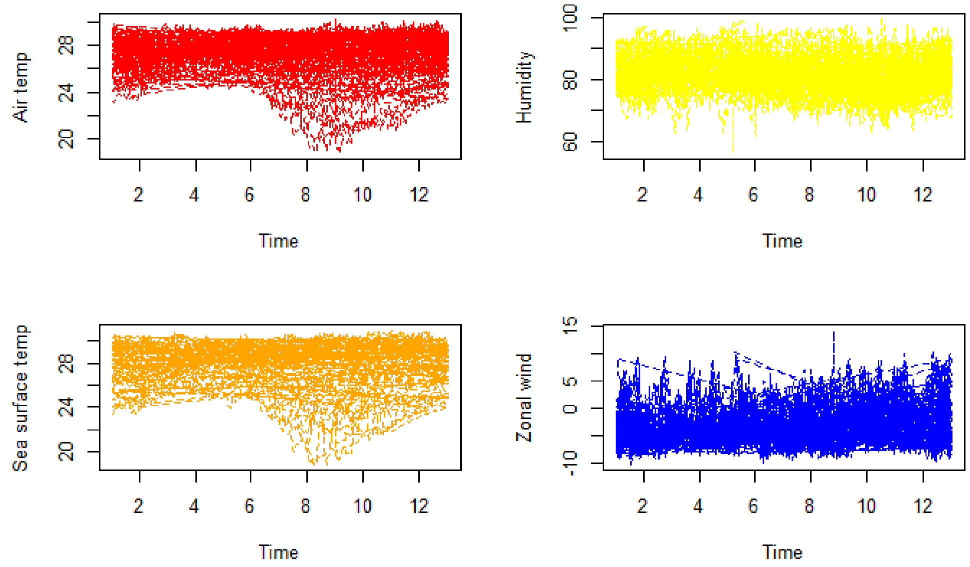
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)         25.7188797  1.0266629  25.051  < 2e-16 ***
zon.winds            1.0218261  0.0458879  22.268  < 2e-16 ***
mer.winds           -0.2228719  0.0476758  -4.675 2.97e-06 ***
humidity            -0.2621548  0.0117482 -22.314  < 2e-16 ***
air.temp             0.0838848  0.0363398   2.308   0.0210 *  
zon.winds:mer.winds -0.0039123  0.0005091  -7.685 1.62e-14 ***
zon.winds:humidity  -0.0005914  0.0002924  -2.022   0.0432 *  
zon.winds:air.temp  -0.0321545  0.0011397 -28.214  < 2e-16 ***
mer.winds:humidity   0.0004656  0.0003401   1.369   0.1710    
mer.winds:air.temp   0.0066521  0.0011035   6.028 1.69e-09 ***
humidity:air.temp    0.0098346  0.0004204  23.395  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5308 on 15750 degrees of freedom
Multiple R-squared:  0.9351,	Adjusted R-squared:  0.9351 
F-statistic: 2.27e+04 on 10 and 15750 DF,  p-value: < 2.2e-16
```

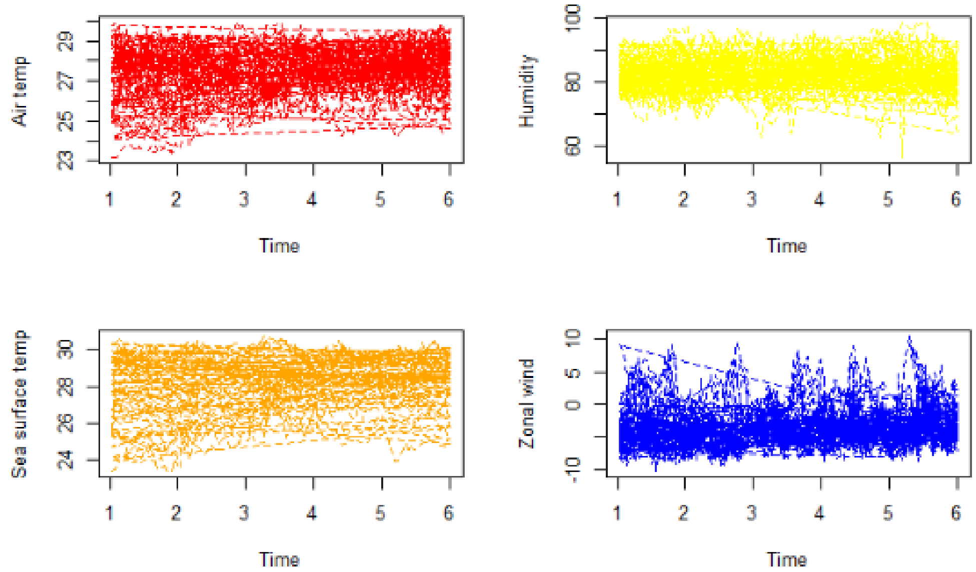
By p-values in the linear regression study, our model concluded that not only four key variables but also higher power and interaction terms are statistically significant. The four measurements collected by the years are critical to predicting sea surface temperatures. Consequently, they should continue to be collected for analysis of sea surface temperature changes.

4. Time Series

To get the time-series plot, first we have to calculate the time for the particular date set by the formula, which is (month + day/30). Here is the plot of time series for the Air temperature, Humidity, Sea Surface temp and Zonal wine. This plot shows the time series for 12 months respectively and we can see that Air temp and Sea surface team is dropping on 8^{th} month and Zonal Wine has high near to 8^{th} month. Apart from this everything is same for whole year. To get more clear view we will plot it for 6 months only.



From the below plot which is for 6months, we can see that the Air temp and Sea Surface temp tends to be high at the start of the year. Also, Zona winds are tending to be higher at the end of each year and Humidity tends to high in the middle of the 3-4 month and 5-6 month.



Code:

```
```{r}
elnino = read.csv("tao-all2.dat.gz", as.is = FALSE, sep = "", header = FALSE, na.strings=".")
elnino_colm = read.csv("tao-all2.col", as.is = FALSE, sep = "", header = FALSE)
names(elnino) <- c("obs","year","month","day","date","latitude","longtitude","zon.winds","mer.winds","humidity","air.temp","ss.temp")


elnino = na.omit(elnino)

summary(elnino)

```
```{r}
year94 = elnino[elnino$year == 94,]
year94 = subset(year94,select = -c(obs,date))

summary(year94)
head(year94)
```
```{r}
library(corrplot)
str(year94)

E2 = cor(subset(year94,select = -c(year,month,day)), method = 'spearman')
corrplot(E2, method="number")
```
```

```
```{r}
year941 = subset(year94, select = -c(1,2,3,4,5))
model = lm(ss.temp~.*., data = year941)

summary.lm(model)
par(mfrow= c(2,2))
plot(model)
```
```

```
```{r}

#time series plot
par(mfcol=(c(2,2)))
summary(year94)
year94$time <- year94$month + year94$day/30

plot(year94$time, year94$air.temp ,type="l",lty=2, xlab="Time",ylab="Air temp",col="red")
plot(year94$time, year94$ss.temp ,type="l",lty=2, xlab="Time",ylab="Sea surface temp",col="orange")
plot(year94$time, year94$humidity ,type="l",lty=2, xlab="Time",ylab="Humidity",col="yellow")
plot(year94$time, year94$zon.winds ,type="l",lty=2, xlab="Time",ylab="Zonal wind",col="blue")

```
```

```
first.half<- year95[year95$time<=6,]

plot(first.half$time, first.half$air.temp ,type="l",lty=2, xlab="Time",ylab="Air temp",col="red")
plot(first.half$time, first.half$ss.temp ,type="l",lty=2, xlab="Time",ylab="Sea surface temp",col="orange")
plot(first.half$time, first.half$humidity ,type="l",lty=2, xlab="Time",ylab="Humidity",col="yellow")
plot(first.half$time, first.half$zon.winds ,type="l",lty=2, xlab="Time",ylab="Zonal wind",col="blue")


```
```
```

Student Performance

- ### Introduction

The dataset chosen for the analysis was the **Student Performance**. Because it has 100 instances and more than ten attributes in the dataset, I have chosen this dataset. Aside from that, Due to the fact that G3 is the final year grade (issued at the 3rd period), while G1 and G2 are the grades for the 1st and 2nd periods, respectively. In the absence of G2 and G1, only G3 can be predicted, but such a forecast is much more useful. All the observations observed during the analysis are included in the report. In order for the reader to get the most out of the dataset, we have included all necessary plots and graphs. A complete analysis was conducted using R studio using the R programming language.

- ### About Dataset

Based on data from two Portuguese schools, this study analyses student achievement in secondary education. The data attributes include student grades, demographics, social characteristics and school-related characteristics, which were collected by school reports and questionnaires. Presented here are two datasets on the performance in two different subjects: Mathematics (mat) and Portuguese (por). A note of importance: the target attribute G3 correlates strongly with attributes G2 and G1.

Attribute Information:

Using R, Both the dataset imported in RStudio, it appeared to be loaded correctly. Then, I have merged both dataset in one data frame by its attributes. There are 395 observations and 33 variables in Math data frame and 649 observation and 33 variables in Portuguese data frame. I combined both data frame and added 2 new columns containing the average score and subject column to determine the main dataset. So, the combine newly formed dataset contains 1044 observations and 35 variables.

Here is the summarized dataset of both dataset:

```
 school   sex          age       address famsize   Pstatus      Medu            Fedu             Mjob          Fjob            reason   
 GP:349   F:208   Min.   :15.0   R: 88   GT3:281   A: 41   Min.   :0.000   Min.   :0.000   at_home : 59   at_home : 20   course    :145  
 MS: 46   M:187   1st Qu.:16.0   U:307   LE3:114   T:354   1st Qu.:2.000   1st Qu.:2.000   health  : 34   health  : 18   home      :109  
                  Median :17.0                             Median :3.000   Median :2.000   other   :141   other   :217   other     : 36  
                  Mean   :16.7                             Mean   :2.749   Mean   :2.522   services:103   services:111   reputation:105  
                  3rd Qu.:18.0                             3rd Qu.:4.000   3rd Qu.:3.000   teacher : 58   teacher : 29                   
                  Max.   :22.0                             Max.   :4.000   Max.   :4.000                                                 
  guardian     traveltime      studytime       failures      schoolsup famsup     paid     activities nursery    higher    internet  romantic 
 father: 90   Min.   :1.000   Min.   :1.000   Min.   :0.0000   no :344   no :153   no :214   no :194    no : 81   no : 20   no : 66   no :263  
 mother:273   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   yes: 51   yes:242   yes:181   yes:201    yes:314   yes:375   yes:329   yes:132  
 other : 32   Median :1.000   Median :2.000   Median :0.0000                                                                                 
              Mean   :1.448   Mean   :2.035   Mean   :0.3342                                                                                 
              3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000                                                                                 
              Max.   :4.000   Max.   :4.000   Max.   :3.0000                                                                                 
     famrel         freetime         goout            Dalc            Walc           health         absences           G1              G2       
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   : 0.000   Min.   : 3.00   Min.   : 0.00  
 1st Qu.:4.000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:3.000   1st Qu.: 0.000   1st Qu.: 8.00   1st Qu.: 9.00  
 Median :4.000   Median :3.000   Median :3.000   Median :1.000   Median :2.000   Median :4.000   Median : 4.000   Median :11.00   Median :11.00  
 Mean   :3.944   Mean   :3.235   Mean   :3.109   Mean   :1.481   Mean   :2.291   Mean   :3.554   Mean   : 5.709   Mean   :10.91   Mean   :10.71  
 3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:5.000   3rd Qu.: 8.000   3rd Qu.:13.00   3rd Qu.:13.00  
 Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :75.000   Max.   :19.00   Max.   :19.00  
       G3       
 Min.   : 0.00  
 1st Qu.: 8.00  
 Median :11.00  
 Mean   :10.42  
 3rd Qu.:14.00  
 Max.   :20.00  
```

```
 school   sex          age        address famsize   Pstatus      Medu            Fedu             Mjob          Fjob            reason   
 GP:423   F:383   Min.   :15.00   R:197   GT3:457   A: 80   Min.   :0.000   Min.   :0.000   at_home :135   at_home : 42   course    :285  
 MS:226   M:266   1st Qu.:16.00   U:452   LE3:192   T:569   1st Qu.:2.000   1st Qu.:1.000   health  : 48   health  : 23   home      :149  
                  Median :17.00                             Median :2.000   Median :2.000   other   :258   other   :367   other     : 72  
                  Mean   :16.74                             Mean   :2.515   Mean   :2.307   services:136   services:181   reputation:143  
                  3rd Qu.:18.00                             3rd Qu.:4.000   3rd Qu.:3.000   teacher : 72   teacher : 36                   
                  Max.   :22.00                             Max.   :4.000   Max.   :4.000                                                 
  guardian     traveltime      studytime       failures      schoolsup famsup     paid     activities nursery    higher    internet  romantic 
 father:153   Min.   :1.000   Min.   :1.000   Min.   :0.0000   no :581   no :251   no :610   no :334    no :128   no : 69   no :151   no :410  
 mother:455   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   yes: 68   yes:398   yes: 39   yes:315    yes:521   yes:580   yes:498   yes:239  
 other : 41   Median :1.000   Median :2.000   Median :0.0000                                                                                 
              Mean   :1.569   Mean   :1.931   Mean   :0.2219                                                                                 
              3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000                                                                                 
              Max.   :4.000   Max.   :4.000   Max.   :3.0000                                                                                 
     famrel         freetime        goout            Dalc            Walc          health         absences            G1             G2       
 Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   : 0.000   Min.   : 0.0   Min.   : 0.00  
 1st Qu.:4.000   1st Qu.:3.00   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.00   1st Qu.:2.000   1st Qu.: 0.000   1st Qu.:10.0   1st Qu.:10.00  
 Median :4.000   Median :3.00   Median :3.000   Median :1.000   Median :2.00   Median :4.000   Median : 2.000   Median :11.0   Median :11.00  
 Mean   :3.931   Mean   :3.18   Mean   :3.185   Mean   :1.502   Mean   :2.28   Mean   :3.536   Mean   : 3.659   Mean   :11.4   Mean   :11.57  
 3rd Qu.:5.000   3rd Qu.:4.00   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.00   3rd Qu.:5.000   3rd Qu.: 6.000   3rd Qu.:13.0   3rd Qu.:13.00  
 Max.   :5.000   Max.   :5.00   Max.   :5.000   Max.   :5.000   Max.   :5.00   Max.   :5.000   Max.   :32.000   Max.   :19.0   Max.   :19.00  
       G3       
 Min.   : 0.00  
 1st Qu.:10.00  
 Median :12.00  
 Mean   :11.91  
 3rd Qu.:14.00  
 Max.   :19.00  
```

- **Data Analysis**

From the summary, it shows that the average grade of student 11.27 and we can also see that the age of students goes from 15 to 22 for this dataset. I have checked for whole the dataset there is no NA values in this dataset.

Here is the first 6 observations of the dataset:

school <fctr>	sex <fctr>	age <int>	address <fctr>	famsize <fctr>	Pstatus <fctr>	Medu <int>	Fedu <int>	Mjob <fctr>	Fjob <fctr>
GP	F	18	U	GT3	A	4	4	at_home	teacher
GP	F	17	U	GT3	T	1	1	at_home	other
GP	F	15	U	LE3	T	1	1	at_home	other
GP	F	15	U	GT3	T	4	2	health	services
GP	F	16	U	GT3	T	3	3	other	other
GP	M	16	U	LE3	T	4	3	services	other

reason <fctr>	guardian <fctr>	traveltime <int>	studytime <int>	failures <int>	schoolsup <fctr>	famsup <fctr>	paid <fctr>	activities <fctr>	nursery <fctr>
course	mother	2	2	0	yes	no	no	no	yes
course	father	1	2	0	no	yes	no	no	no
other	mother	1	2	3	yes	no	yes	no	yes
home	mother	1	3	0	no	yes	yes	yes	yes
home	father	1	2	0	no	yes	yes	no	yes
reputation	mother	1	2	0	no	yes	yes	yes	yes

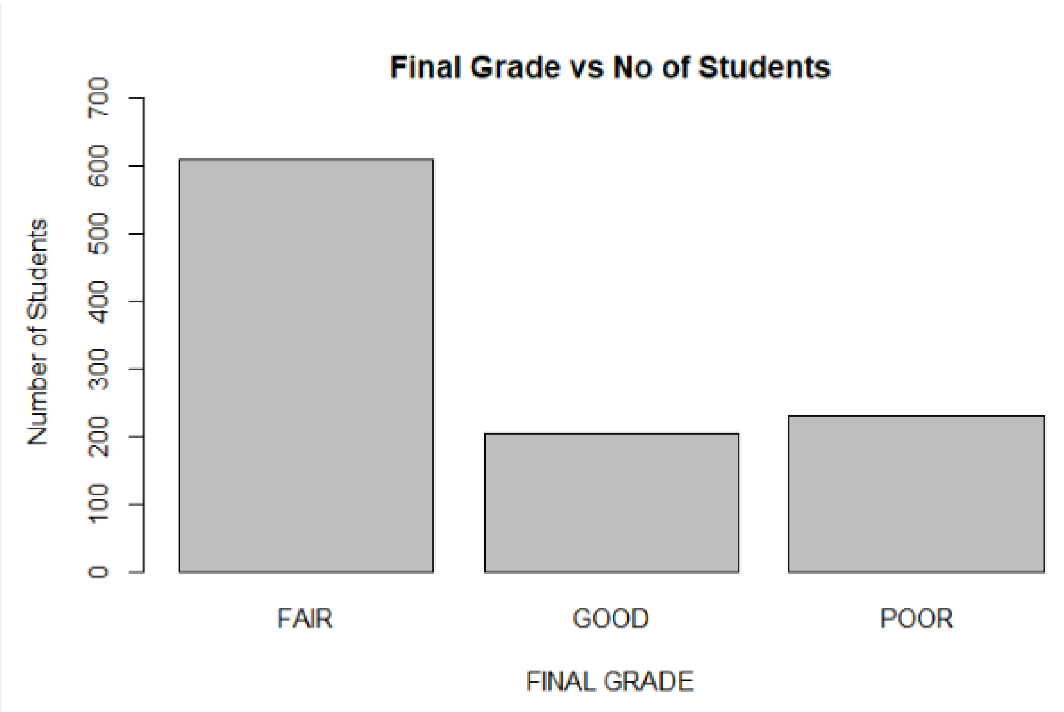
higher <fctr>	internet <fctr>	romantic <fctr>	famrel <int>	freetime <int>	goout <int>	Dalc <int>	Walc <int>	health <int>	absences <int>
yes	no	no	4	3	4	1	1	3	6
yes	yes	no	5	3	3	1	1	3	4
yes	yes	no	4	3	2	2	3	3	10
yes	yes	yes	3	2	2	1	1	5	2
yes	no	no	4	3	2	1	2	5	4
yes	yes	no	5	4	2	1	2	5	10

G1 <int>	G2 <int>	G3 <int>	average <dbl>	subject <chr>
5	6	6	5.67	Math
5	5	6	5.33	Math
7	8	10	8.33	Math
15	14	15	14.67	Math
6	10	10	8.67	Math
15	15	15	15.00	Math

- **Data Visualisation**

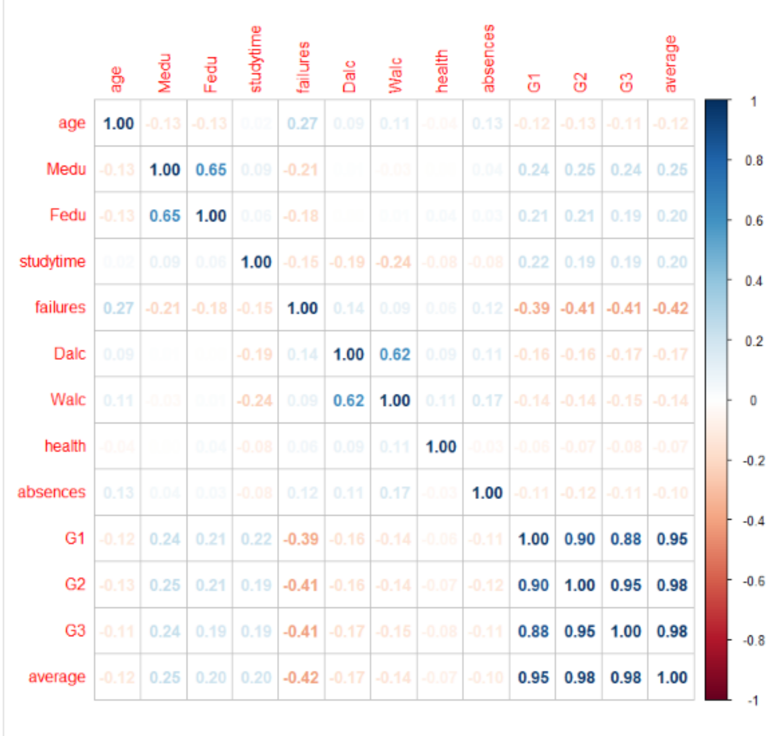
1. **Histogram**

I added one more categorical column in the combine dataset which is the final grade as Good for the grade of greater than equal to 15 and less than equal to 20, Fair for the grade of greater than equal to 10 and less than equal to 14 and Poor for the grade of greater than equal to 0 and less than equal to 9 accordingly. Below is the histogram for the same. From the below Histogram we can see that the Number of students for Fair grade are more than other categories.



2. Correlation Plot

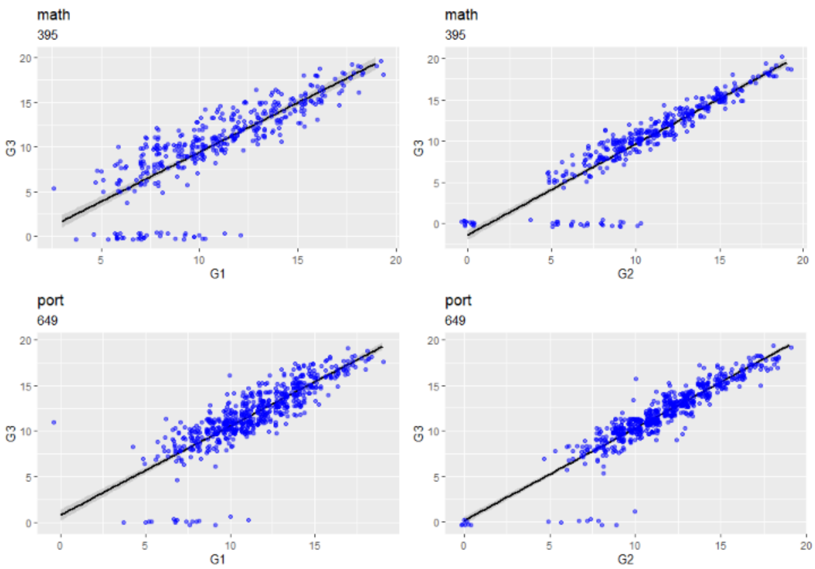
I. Correlation plot for Both Subject



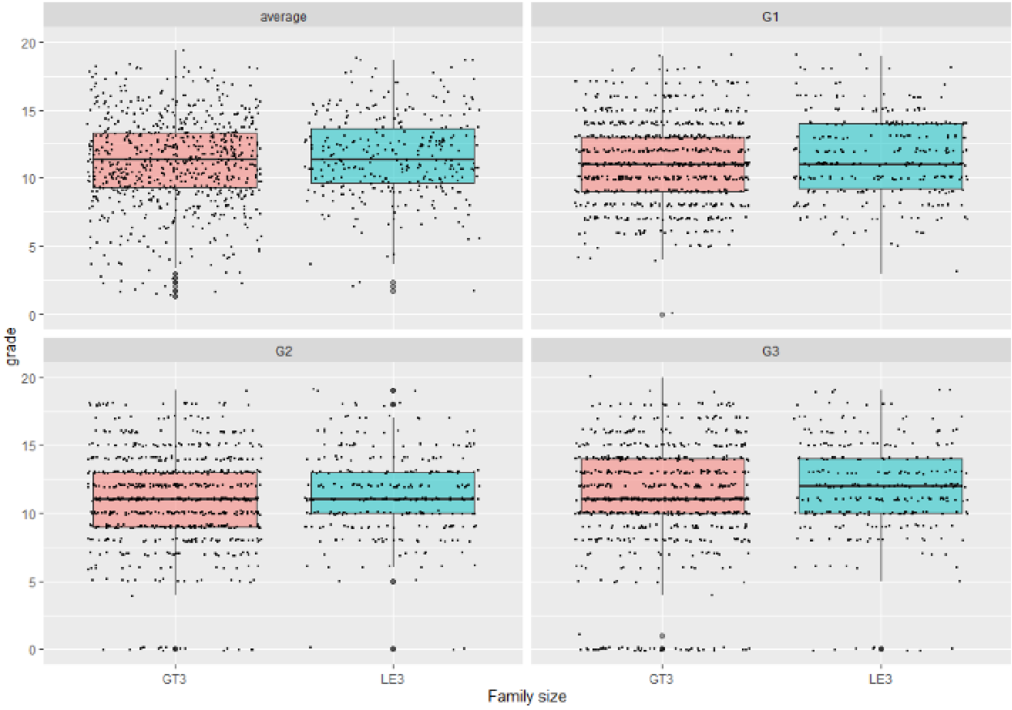
From both Correlation plot we can see that no specific variable is highly positive or highly negative correlated with grade. However, as describe earlier that grade shows highly correlation in between them. So, we have to check this with the different method keeping all the variable.

3. Box plot:

Moving further, I have taken two datasets: one for the children whose families live with them and another for the children whose families do not live with them. This plot shows if there is a correlation between the three exams for both datasets for two different subjects after splitting.



Moving further to the part, now I will consider the family dataset. Let's see if family size and parental status affects a student's grade




```
        Two Sample t-test

data:  average by famsize
t = -2.0017, df = 1042, p-value = 0.9772
alternative hypothesis: true difference in means between group GT3 and group LE3 is greater than 0
95 percent confidence interval:
 -0.7972451        Inf
sample estimates:
mean in group GT3 mean in group LE3
         11.13898          11.57644


        Two Sample t-test

data:  average by Pstatus
t = 0.5107, df = 1042, p-value = 0.3048
alternative hypothesis: true difference in means between group A and group T is greater than 0
95 percent confidence interval:
 -0.3535435        Inf
sample estimates:
mean in group A mean in group T
       11.40777        11.24878
```

On the basis of the p-value, let’s see if the parent status affects the student scores. We can see that; The difference is not significant.

Children whose parents live apart and those whose parents live together have similar average grades. It is possible to draw the same conclusion about students living in families smaller than or equal to 3 people and those living with more than 3 people. Thus, parental status and family size have no significant impact on grades.

Quite a few families with parents living together have students with zero marks on the final exam, in comparison with families with separated parents. Similar trends are observed in families with more than 3 members in comparison to families with up to 3 members.

According to the graphs above, we can also notice that students have more difficulty in the second period than in the first, which is not surprising, since the difficulty level would most likely be higher towards the end of the year. To conclude we can see that the on the basis of p-value we can say that, family size and parental status don't contribute much more. So, we will move to the next variables.

```
Call:
lm(formula = average ~ Mjob, data = family)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.4454  -1.8379   0.0368   2.1108   7.8901 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    10.4399     0.2271  45.968  < 2e-16 ***
Mjobhealth      2.0680     0.4167   4.963 8.10e-07 ***
Mjobother       0.5233     0.2769   1.890 0.059051 .  
Mjobservices    1.1192     0.3057   3.661 0.000264 ***
Mjobteacher     1.6754     0.3585   4.673 3.36e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.163 on 1039 degrees of freedom
Multiple R-squared:  0.03792,	Adjusted R-squared:  0.03421 
F-statistic: 10.24 on 4 and 1039 DF,  p-value: 3.932e-08
```

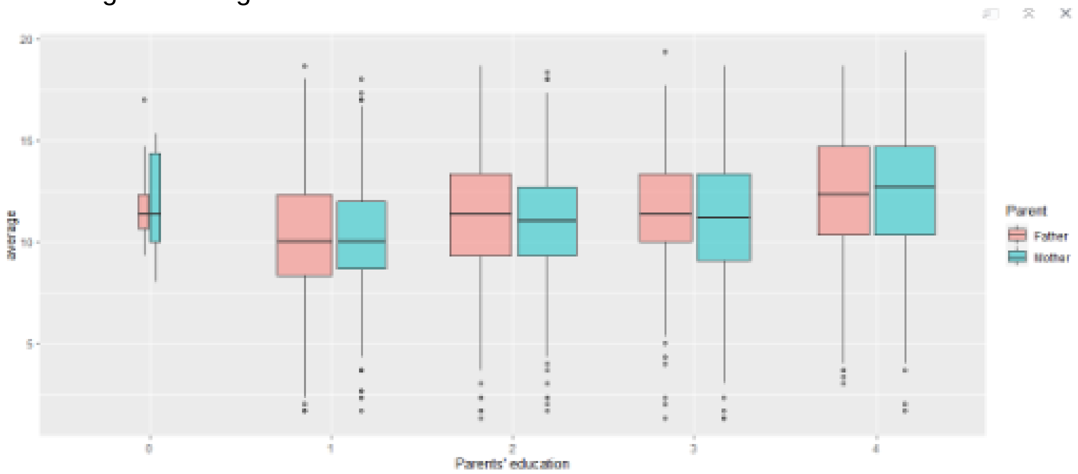
```
Call:
lm(formula = average ~ Fjob, data = family)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.8615 -1.8193 -0.1416  2.1807  8.1884 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    10.8553     0.4056  26.761  < 2e-16 ***
Fjobhealth      1.0800     0.6429   1.680 0.093287 .  
Fjobother       0.2940     0.4266   0.689 0.490864    
Fjobservices    0.2863     0.4466   0.641 0.521624    
Fjobteacher     2.0062     0.5670   3.538 0.000421 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.194 on 1039 degrees of freedom
Multiple R-squared:  0.01913,	Adjusted R-squared:  0.01536 
F-statistic: 5.067 on 4 and 1039 DF,  p-value: 0.0004791
```

The scores of students whose mothers work in the health industry tend to be higher than most of the others, and the scores of students whose mothers are *housewives* tend to be lower. There is a difference in average scores according to the jobs of the fathers. Students with a father who is a teacher have a higher average score overall.




```
Call:
lm(formula = average ~ Medu, data = family)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.5040  -1.8440   0.1246   2.1246   7.7739 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  9.57679    0.24483  39.117  < 2e-16 ***
Medu         0.64930    0.08633   7.521 1.17e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.136 on 1042 degrees of freedom
Multiple R-squared:  0.05149,	Adjusted R-squared:  0.05058 
F-statistic: 56.56 on 1 and 1042 DF,  p-value: 1.171e-13
```

```
Call:
lm(formula = average ~ Fedu, data = family)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.2714  -1.8175   0.0686   2.0686   8.1607 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  9.96326    0.23414  42.553  < 2e-16 ***
Fedu         0.54606    0.08906   6.131 1.24e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.164 on 1042 degrees of freedom
Multiple R-squared:  0.03482,	Adjusted R-squared:  0.03389 
F-statistic: 37.59 on 1 and 1042 DF,  p-value: 1.237e-09
```

Parents with primary school education score significantly lower than others on average, while parents with higher education score significantly higher. The large P-value indicates that family relationships do not have much to do with student performance.

Conclusions show that parents' education and jobs are the two most important factors affecting a student's overall performance. Among these two factors, education factor confirms the common notion that children will benefit from parents with higher education degrees. Furthermore, there is interesting information regarding the job factor that indicates that the students with a mother who works in the health industry and a father who works as a teacher tend to perform better overall.

Code Chunk :

```
data <- rbind(math, port)
str(data)
summary(data)
head(data)
```

```
```{r}
final_grade_math = math
good = math[((math$G3>=15) & (math$G3<= 20))]
fair = math[((math$G3>=10) & (math$G3<= 14))]
poor = math[((math$G3>=0) & (math$G3<= 9))]
good$grade = c("GOOD")
fair$grade = c("FAIR")
poor$grade = c("POOR")

final_grade_math = list(good,poor, fair)
library(reshape)
final_grade_math = (merge_recurse(final_grade_math))


final_grade_port = port
goodp = port[((port$G3>=15) & (port$G3<= 20))]
fairp = port[((port$G3>=10) & (port$G3<= 14))]
poorp = port[((port$G3>=0) & (port$G3<= 9))]
goodp$grade = c("GOOD")
fairp$grade = c("FAIR")
poorp$grade = c("POOR")

final_grade_port = list(goodp,poorp, fairp)

final_grade_port = (merge_recurse(final_grade_port))
final_grade_math$grade =  as.factor(final_grade_math$grade)
final_grade_port$grade =  as.factor(final_grade_port$grade)
```

```
family %>%
  select(Mjob, Fjob, average)%>%
  gather(key=Parent, value=job, -average) %>%
  ggplot(aes(job, average, fill=Parent))+
  geom_boxplot(varwidth=T, alpha=0.5)+
  xlab("Parents' jobs")+
  scale_fill_discrete(labels=c("Father","Mother"))
```
```{r}

model1 = lm(average~Mjob, family)
summary(model1)

```

```{r}

model2 = lm(average~Fjob, family)
summary(model2)

```
```

```
family %>%
  select(Medu, Fedu, average) %>%
  gather(key=Parent, value=education, -average) %>%
  ggplot(aes(x= factor(education), y= average, fill=Parent))+
  geom_boxplot(varwidth=T, alpha=0.5)+
  xlab("Parents' education")+
  scale_fill_discrete(labels=c("Father","Mother"))
```
```{r}
model3 = lm(average~Medu, family)
summary(model3)
```
```{r}
model4 = lm(average~Fedu, family)
summary(model4)
```
```
```

Wine Quality

- ## Introduction

The dataset chosen for the analysis was the **Wine quality**. It was chosen since there are over 100 instances and more than ten attributes in the dataset. Aside from that, I picked this dataset because of its perspective on Wine quality. All the observations observed during the analysis are included in the report. Specifically, we aimed to determine which factors contribute to wine quality by analysing the dataset. Also, This report explores the relationship of wine between the variable quality and its chemical attributes. In order for the reader to get the most out of the dataset, we have included all necessary plots and graphs. A complete analysis was conducted using R studio using the R programming language.

- ## About Dataset

Red and white Portuguese "Vinho Verde" wines are the subjects of the two datasets. In order to protect privacy and logistic issues, only physicochemical (the inputs) and sensory (the output) variables are available. For example, grape types, wine brands, or selling prices of wines are not included in the dataset. Classification is ordered and unbalanced (e.g., there are many more normal wines than excellent or poor wines). A few excellent or poor wines could be detected by using outlier detection algorithms. However, we are not certain that all the input variables are relevant. Thus, it may be worthwhile to test various selection methods. There are 12 attributes and 4898 Instances of white wine and 12 attributes and 1599 observation of red wine.

Attribute Information:

1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar
5 - chlorides
6 - free sulfur dioxide
7 - total sulfur dioxide
8 - density
9 - pH
10 - sulphates
11 - alcohol
Output variable (based on sensory data):
12 - quality (score between 0 and 10)

- ## Data Analysis

Using R, the both datasets were imported into a data frame in csv format. When the dataset was first examined, it appeared to be loaded correctly. Then, I combined the red wine and white wine datasets, and added the code column to data frame where red wine as 1 and white wine as 0. There are 4898 observations of white wine and 1599 observations of red wine each frame with 13 variables giving us a 6497 by 13 data frame.

Here is the summarized dataset after validation:

```
 fixed.acidity    volatile.acidity  citric.acid     residual.sugar     chlorides       free.sulfur.dioxide total.sulfur.dioxide    density             pH
 Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600   Min.   :0.00900   Min.   :  1.00      Min.   :  6.0        Min.   :0.9871   Min.   :2.720
 1st Qu.: 6.400   1st Qu.:0.2300   1st Qu.:0.2500   1st Qu.: 1.800   1st Qu.:0.03800   1st Qu.: 17.00      1st Qu.: 77.0        1st Qu.:0.9923   1st Qu.:3.110
 Median : 7.000   Median :0.2900   Median :0.3100   Median : 3.000   Median :0.04700   Median : 29.00      Median :118.0        Median :0.9949   Median :3.210
 Mean   : 7.215   Mean   :0.3397   Mean   :0.3186   Mean   : 5.443   Mean   :0.05603   Mean   : 30.53      Mean   :115.7        Mean   :0.9947   Mean   :3.219
 3rd Qu.: 7.700   3rd Qu.:0.4000   3rd Qu.:0.3900   3rd Qu.: 8.100   3rd Qu.:0.06500   3rd Qu.: 41.00      3rd Qu.:156.0        3rd Qu.:0.9970   3rd Qu.:3.320
 Max.   :15.900   Max.   :1.5800   Max.   :1.6600   Max.   :65.800   Max.   :0.61100   Max.   :289.00      Max.   :440.0        Max.   :1.0390   Max.   :4.010
   sulphates         alcohol          quality       code
 Min.   :0.2200   Min.   : 8.00   Min.   :3.000   1:1599
 1st Qu.:0.4300   1st Qu.: 9.50   1st Qu.:5.000   0:4898
 Median :0.5100   Median :10.30   Median :6.000
 Mean   :0.5313   Mean   :10.49   Mean   :5.818
 3rd Qu.:0.6000   3rd Qu.:11.30   3rd Qu.:6.000
 Max.   :2.0000   Max.   :14.90   Max.   :9.000
```

From the summary, it shows that there were more white wines than red and they were roughly in 3 to 1 ratio. In addition to these, it can be seen that there are average of 10.49 percentage of alcohol included in the wines. Also, it can be seen that sometimes the alcohol percentage is 14.90 as well. Along with this, pH scale for Wines goes from 0 to 7 where there is average of 3.219 for the wines.

Here is the first 10 observations of the dataset:

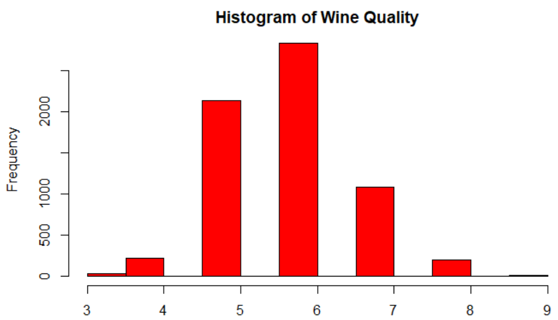
	fixed.acidity <dbl>	volatile.acidity <dbl>	citric.acid <dbl>	residual.sugar <dbl>	chlorides <dbl>	free.sulfur.dioxide <dbl>	total.sulfur.dioxide <dbl>	density <dbl>	pH <dbl>	sulphates <dbl>	alcohol <dbl>	quality <int>	code <dbl>
1	7.4	0.70	0.00	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5	1
2	7.8	0.88	0.00	2.6	0.098	25	67	0.9968	3.20	0.68	9.8	5	1
3	7.8	0.76	0.04	2.3	0.092	15	54	0.9970	3.26	0.65	9.8	5	1
4	11.2	0.28	0.56	1.9	0.075	17	60	0.9980	3.16	0.58	9.8	6	1
5	7.4	0.70	0.00	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5	1
6	7.4	0.66	0.00	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5	1

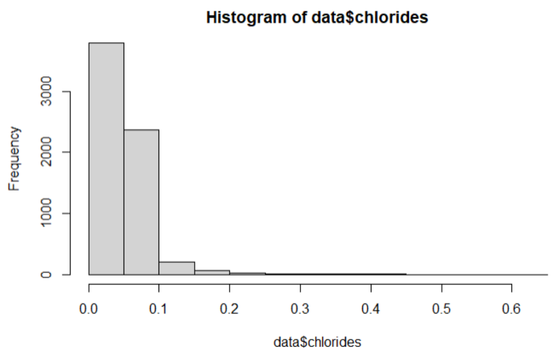
6 rows | 1-10 of 13 columns

- **Data Visualisation**

1. Histogram

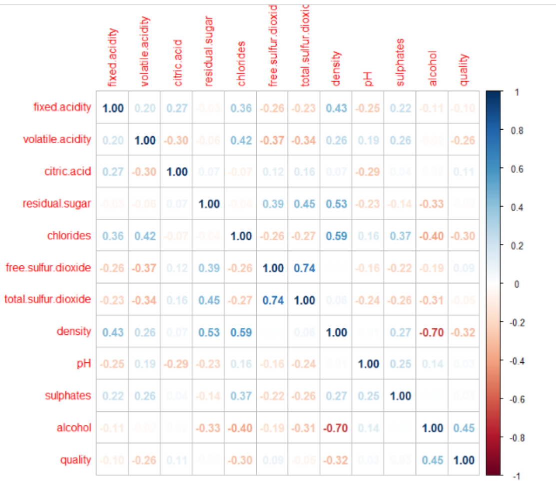
First, I will look at the variable quality. Each expert graded the wine quality between 0 (very bad) and 10 (very excellent). Below, I see that the bulk of the wine quality is at a quality of 5, 6 and 7. There is no observations below a quality of 3 and none above 9. Also, we can see that chloride dataset is not normal so we have to perform the log transform on it. We can see that from the histogram plot as below:

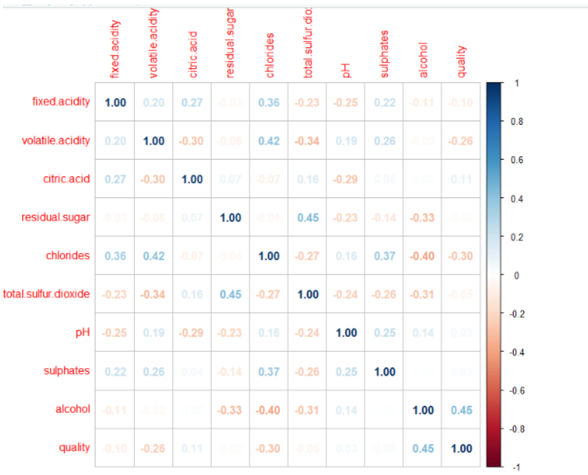




2. Correlation Plot

Now let's see how the continuous variables correlate with each other. The multicollinearity of predictor variables, where one predictor variable is highly correlated with another, is a common concern in data analysis. In addition to making parameter estimation unstable, multicollinearity makes understanding the effect of the predictor difficult. My goal is to identify highly correlated variables and eliminate them from future analyses. I have included two correlation plots below. The first illustrates the relationship between all 13 variables. The second demonstrates the relationship after highly correlated variables are removed.

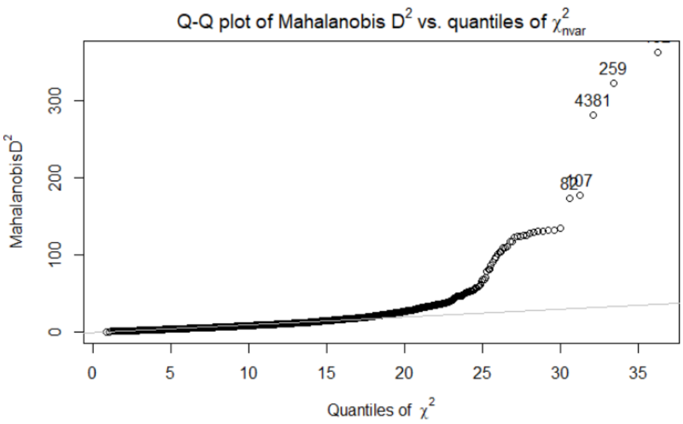




According to the first correlation plot, alcohol and density have a negative linear correlation of 0.7. The positive linear correlation between free sulfur dioxide and total sulfur dioxide is 0.74. Such highly correlated variables will complicate the analysis. Consequently, density and sulfur dioxide will be excluded from the analysis.

3. QQ plot

Next, we need to consider whether there are any outliers in our data. The analysis can be complicated by outliers, as the model(s) may be skewed towards those extreme value(s). We have a function called outliers in the PSYCH library for detecting outliers. This plot illustrates how it works. The last five observations on the plot are extreme values, so I will focus on these five.



Observe five extreme values identified as 152, 259, 4381, 107 , 82 in the dataset.

	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	total.sulfur.dioxide	pH	sulphates	alcohol	quality
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>
152	9.2	0.520	1.00	3.4	0.610	69	2.74	2.00	9.4	4
259	7.7	0.410	0.76	1.8	0.611	45	3.06	1.26	9.4	5
4381	7.8	0.965	0.60	65.8	0.074	160	3.39	0.69	11.7	6
107	7.8	0.410	0.68	1.7	0.467	69	3.08	1.31	9.3	5
82	7.8	0.430	0.70	1.9	0.464	67	3.13	1.28	9.4	5

5 rows | 1-9 of 10 columns

The variable Quality is ranged from 0 to 10 in increments of 1. However, the observations in the data set only go from 3 to 9.

```
   3    4    5    6    7    8    9 
  30  216 2138 2836 1079  193    5 
```

4. Linear Regression

```
Call:
lm(formula = quality ~ volatile.acidity + residual.sugar + chlorides + 
    total.sulfur.dioxide + pH + sulphates + alcohol, data = data4)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.4825 -0.4630 -0.0327  0.4671  3.0209 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          1.6855145  0.2251592   7.486 8.05e-14 ***
volatile.acidity    -1.4464944  0.0673446 -21.479  < 2e-16 ***
residual.sugar       0.0248910  0.0023968  10.385  < 2e-16 ***
chlorides           -0.0631007  0.0292657  -2.156   0.0311 *  
total.sulfur.dioxide -0.0011125  0.0002127  -5.230 1.75e-07 ***
pH                   0.1940223  0.0614379   3.158   0.0016 ** 
sulphates            0.6897831  0.0696345   9.906  < 2e-16 ***
alcohol              0.3277389  0.0096656  33.908  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7397 on 6484 degrees of freedom
Multiple R-squared:  0.2831,	Adjusted R-squared:  0.2823 
F-statistic: 365.8 on 7 and 6484 DF,  p-value: < 2.2e-16
```

Based on the above analysis, we can see that with the help of step regression, it was known that there were six significant factors that contributed to get the best wine quality. The six factors were volatile acidity, residual sugar, sulphur dioxide, pH scale, sulphates and last but not least alcohol percentage with the p-value less than 0.05. It stated that lower level of volatile acidity, residual sugar, as well as higher levels of sulphur dioxide, pH, and alcohol result in better wine tasting.

On the basis of this, we can see that the best quality wines have high values of both alcohol percentage and Sulphates concentration, so the higher the contents the better the wine quality. It can be treated as the when the concentration of alcohol is higher, Sulphates tend to be lower in good quality wine. For the bad quality wine, the alcohol and Sulphates level are relatively lower than other higher quality wine, and the reduction of percentage of alcohol level is more significant than Sulphates.

Code Chunk:

```
red = read.csv('winequality-red.csv', header = TRUE, sep = ";", as.is = FALSE)
white = read.csv('winequality-white.csv', header = TRUE, sep = ";", as.is = FALSE)
red$code = 1
white$code= 0
red$code = as.factor(red$code)
white$code = as.factor(white$code)

df = list(red,white)
library(reshape)
data = merge_recurse(df)
summary(data)
head(data)

```
```{r}
hist(data$quality, col = 'red', ylab = 'Frequency', xlab = 'Quality', main = 'Histogram of Wine Quality')
```
```

```
```{r, fig.width=8, fig.height=8}
data1 = data
library(corrplot)

#Removing Code as it is not neccessary
data2 = subset(data, select = c(-13))
M1 = cor(data2, method = 'spearman')
corrplot(M1, method="number")

#After removing highly co-related variables
data3 = subset(data2, select = c(-6,-8))
M2 = cor(data3, method = 'spearman')
corrplot(M2, method="number")
str(data3)
```
```

```
```{r}

```
```

```
```{r}
library(psych)
a = outlier(data3, plot = TRUE)
data4 = data3[-c(152, 259, 4381, 107, 82), ]
data3[c(152, 259, 4381, 107, 82), ]
```
```

```
```{r}

model = lm(quality~., data = data4)
smodel = step(model,trace = FALSE)
summary(smodel)
par(mfrow= c(2,2))
plot(smodel)
```
```
```